RCnum: A Semantic and Multilingual Online Edition of the Geneva Council Registers from 1545 to 1550

Pierrette Bouillon[1], **Christophe Chazalon**[2], **Sandra Coram-Mekkey**[3], **Gilles Falquet**[2], **Johanna Gerlach**[1], **Stéphane Marchand-Maillet**[2], **Laurent Moccozet**[2], **Jonathan Mutal**[1], **Raphael Rubino**[1] and **Marco Sorbi**[2]

[1] TIM/FTI, University of Geneva, 1205 Geneva – Switzerland
{firstName.lastName}@unige.ch
[2] CUI, University of Geneva, 1205 Geneva – Switzerland
{firstName.lastName}@unige.ch
[3] Fondation de l'Encyclopédie de Genève
coram.mekkey@gmail.com

Abstract

The RCnum project is funded by the Swiss National Science Foundation and aims at producing a multilingual and semantically rich online edition of the Registers of Geneva Council from 1545 to 1550. Combining multilingual NLP, history and paleography, this collaborative project will clear hurdles inherent to texts manually written in 16th century Middle French while allowing for easy access and interactive consultation of these archives.

1 Introduction

The RCnum[1] project aims at producing a semantic and multilingual online edition of the Geneva Council Registers (*Registres du Conseil de Genève*, RC hereafter) for the years 1545 to 1550. This project, which began in July 2023 and will run until June 2027, is based on a synergy between the Fondation de l'Encyclopédie de Genève and two Geneva University faculties, namely the Centre universitaire d'informatique (CUI) and the Faculty of translation and interpreting (FTI). Previous work on RC have focused on manual transcription and editing, leading to in print publication for the years 1536 to 1544. Manual transcription from 1545 to 1550 has been conducted prior to this project, resulting in a digitised version of this corpus. RCnum's objective is to continue the digitisation effort while automatising RC modernisation, and to develop new functionalities to make RC accessible to a wide audience. RCnum is divided into four work packages (WP): 1) RC collection and preparation following Text Encoding Initiative guidelines, 2) automatic normalisation, modernisation and translation, 3) development of interactive and pedagogical exploration and visualisation tools, and 4) indexing, semantic enrichment and online platform development.

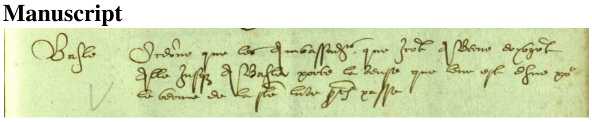
2 Project Description

Overview Comprising an uninterrupted series from 1409 until today, the RC document the work of the Geneva authorities, providing insights into society and politics over time. The RC in the time of Calvin, in particular, is an invaluable resource for the study of the political, legal, economic, social, and religious history of the Geneva region. These archives are also of philological interest, since they use Middle French for judicial and administrative matters which had until then been written in Latin. However, these documents remain difficult to understand for non-historians, paleographers, or experts in Middle French with knowledge of the political landscape during this period. Previous work in French modernisation has explored the use of machine translation (MT) approaches, including neural MT (e.g. Transformer (Vaswani et al., 2017)). Yet, these methods require large parallel corpora which are lacking for the RC content in Middle French.

Expected Outcomes The main outcome of RCnum is an open-source ergonomic and dynamic platform which will offer RC original transcriptions along with their normalised, modernised, translated and semantically enriched content, accessible through data exploration tools. For the normalisation and modernisation tasks (detailed in Figure 1), we will leverage low-resource MT techniques, e.g. fine-tuning large language models (LLMs) and artificial data generation, resulting

[1] https://www.unige.ch/registresconseilge

Manuscript



Step 1: Manual transcription
(Basle) — Ordonne que les ambassadeurs que iront a Berne doybgent alle jusque a Basle porte la cense que leur est dhue pour le terme de la saincte luce prochain passe.

Step 2: Local normalisation
(Bâle) – Ordonné que les ambassadeurs que iront à Berne doivent aller jusqu'à Bâle porter le cens que leur est dû pour le terme de la Sainte-Luce prochaine passée.

Step 3: Syntactic normalisation
(Bâle) – Il a été ordonné que les ambassadeurs qui iront à Berne doivent aller jusqu'à Bâle porter le cens qui leur est dû pour le terme de la Sainte-Lucie prochaine passée.

Step 4: Modernisation
(Bâle) – Il a été ordonné que les ambassadeurs qui iront à Berne doivent aller jusqu'à Bâle pour apporter les intérêts échus à la Sainte-Lucie passée.

Step 5: Translation
(Basel) – It has been ordered that the ambassadors who will go to Bern must go to Basel to bring the loan interests which were due on the past Saint Lucy's Day.

Figure 1: Sample taken from the Geneva Council meeting minutes held on January 5th, 1545, manually transcribed, normalised, modernised and translated.

in several versions of the corpus linked through token alignments and enriched with external information. Enrichment will be based on representation structures such as knowledge graphs combined with Linked Open Data sources (Hogan et al., 2021; Munnelly et al., 2018). The enriched corpus will contain information about named entities, dates, etc., facilitating RC modernisation with MT techniques enhanced with glossaries covering historical word forms (Dougal and Lonsdale, 2020). Simultaneously to these tasks, we will work on identifying historians and non-experts' needs in terms of user interfaces in order to design an interactive platform based on semantically enriched data visualisation techniques (Knabben et al., 2021). Furthermore, the platform will allow data enrichment through input and validation by the community. Interactive tools adapted to RC content will be evaluated with user-based tests and iterative processes aiming at improving the UI/UX (Isenberg et al., 2013).

First Results RC normalisation experiments are presented in Rubino et al. (2024b), focusing on spelling variants reduction while preserving 16th century historical wordforms. A pre-trained LLM baseline fine-tuned on a small parallel corpus outperformed previously released models trained for the normalisation of Early Modern French, as indicated by automatic metrics. Further experiments with synthetic data generation improved over this baseline. To validate these findings, we conducted a manual evaluation through post-editing, comparing normalisation from scratch to our automatic normalisation approaches (Rubino et al., 2024a).

Acknowledgements

This project is funded by the Swiss National Science Foundation (Grant n. 215733).

References

Dougal, Duane K. and Deryle Lonsdale. 2020. Improving NMT Quality Using Terminology Injection. In *LREC*, pages 4820–4827.

Hogan, Aidan, Eva Blomqvist, Michael Cochez, Claudia D'amato, Gerard De Melo, Claudio Gutierrez, Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, Axel-Cyrille Ngonga Ngomo, Axel Polleres, Sabbir M. Rashid, Anisa Rula, Lukas Schmelzeisen, Juan Sequeda, Steffen Staab, and Antoine Zimmermann. 2021. Knowledge Graphs. *ACM Computing Surveys*, 54(4):1–37.

Isenberg, Tobias, Petra Isenberg, Jian Chen, Michael Sedlmair, and Torsten Möller. 2013. A Systematic Review on the Practice of Evaluating Visualization. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2818–2827.

Knabben, Moritz, Martin Baumann, Tanja Blascheck, Thomas Ertl, and Steffen Koch. 2021. Visualizing Temporal-Thematic Patterns in Text Collections. In *Vision, Modeling, and Visualization*, pages 9–16.

Munnelly, Gary, Harshvardhan J Pandit, and Séamus Lawless. 2018. Exploring Linked Data for the Automatic Enrichment of Historical Archives. In *The Semantic Web: ESWC*, pages 423–433.

Rubino, Raphael, Sandra Coram-Mekkey, Johanna Gerlach, Jonathan Mutal, and Pierrette Bouillon. 2024a. Automatic Normalisation of Middle French and its Impact on Productivity. In *LT4HALA*.

Rubino, Raphael, Johanna Gerlach, Jonathan David Mutal, and Pierrette Bouillon. 2024b. Normalizing without Modernizing: Keeping Historical Wordforms of Middle French while Reducing Spelling Variants. In *Findings of NAACL*.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30.